JOURNAL OF APPLIED STATISTICS
2020, VOL. 47, NO. 12, 2208–2229
https://doi.org/10.1080/02664763.2019.1707517

Taylor & Francis
Taylor & Francis Group

Quasi-binomial zero-inflated regression model suitable for variables with bounded support

E. Gómez–Déniz[a], D. I. Gallardo[b] and H. W. Gómez[c]

[a]Department of Quantitative Methods in Economics and TiDES Institute, University of Las Palmas de Gran Canaria, Las Palmas, Spain; [b]Departamento de Matemática, Universidad de Atacama, Copiapo, Chile; [c]Departamento de Matemáticas, Facultad de Ciencias Básicas, Universidad de Antofagasta, Antofagasta, Chile

ABSTRACT

In recent years, a variety of regression models, including zero-inflated and hurdle versions, have been proposed to explain the case of a dependent variable with respect to exogenous covariates. Apart from the classical Poisson, negative binomial and generalised Poisson distributions, many proposals have appeared in the statistical literature, perhaps in response to the new possibilities offered by advanced software that now enables researchers to implement numerous special functions in a relatively simple way. However, we believe that a significant research gap remains, since very little attention has been paid to the quasi-binomial distribution, which was first proposed over fifty years ago. We believe this distribution might constitute a valid alternative to existing regression models, in situations in which the variable has bounded support. Therefore, in this paper we present a zero-inflated regression model based on the quasi-binomial distribution, taking into account the moments and maximum likelihood estimators, and perform a score test to compare the zero-inflated quasi-binomial distribution with the zero-inflated binomial distribution, and the zero-inflated model with the homogeneous model (the model in which covariates are not considered). This analysis is illustrated with two data sets that are well known in the statistical literature and which contain a large number of zeros.

ARTICLE HISTORY
Received 9 May 2019
Accepted 13 December 2019

KEYWORDS
Fit; quasi binomial distribution; score test; zero-inflated

MATHEMATICS SUBJECT CLASSIFICATIONS (2000)
62F03; 62A05

1. Introduction

In recent years, many zero-inflated models have been proposed to describe data characterised by a preponderance of zeros, perhaps in response to the new possibilities offered by advanced software that now enables researchers to implement numerous special functions in a relatively simple way. These models include the zero-inflated Poisson distribution ([2,9,17,24,26,27]; among others), negative binomial distributions, which are the most commonly used, [27] and recently the generalised Poisson distribution [20,22]. These models are being used increasingly and they are commonly found in numerous fields, in research areas such as manufacturing [26], dental epidemiology [2], domestic violence [20], medicine [9,17] recreational trips [23] and traffic accidents [28]. For a review of the Poisson and negative binomial regression models, see Cameron and Trivedi [8]. All of the

CONTACT E. Gómez–Déniz emilio.gomez-deniz@ulpgc.es

above approaches take into account that the researcher may wish to introduce covariates into the model. Thus, the existing catalogue in this respect is sufficiently extensive to enable a variety of data types (for example, presenting overdispersion or underdispersion) to be modelled. However, none of the above 'classical' models provides the option of incorporating the bounded support variable for the variable under study, with the sole exception of the binomial distribution, and this lacks the necessary flexibility to model data corresponding to different situations. Thus, in order to work with any of the above-mentioned distributions, it is usually necessary to censor the support of the variable.

The quasi binomial distribution, which was first proposed over fifty years ago, can be considered a useful alternative to existing regression models when the support of the variable is bounded [13,30]. Its formulation provides a novel parameter, other than those available with the classical binomial distribution and, therefore, offers the possibility of showing overdispersion (variance larger than the mean) or underdispersion (variance lower than the mean). Despite these advantages, however, little previous research attention has been paid to this possibility.

In our research work, cases have arisen in which the classical Poisson, negative binomial and generalised Poisson distributions have been considered inadequate because they could not be fitted to the data sets. When the variable under study is bounded, a model providing this feature is required. The binomial distribution might be appropriate for this purpose, but unfortunately this is not so when the empirical data present overdispersion. The following examples may be considered of cases in which the values of the study variable are bounded: in surveys of consumer activity, the practitioner often requires respondents to report their actual behaviour over a relatively short period of past time, the reference period usually being taken as the last seven days [1]; in dental epidemiological analysis, the practitioner sometimes analyses count data such as caries, a variable that is bounded by the maximum number of teeth; in the study of health services that we consider in this paper, one of the variables of interest is the number of days of reduced activity during the past two weeks, and so the maximum value of this variable is 14.

In this paper, we consider a homogeneous (the model in which covariates are not considered), zero-inflated regression model based on the quasi-binomial distribution introduced by Consul [12]. Here, the binomial distribution is a particular case and counts that are bounded between zero and a given number can be considered. When a data set containing a large number of zeros must be analysed, a possible approach is to use a zero-inflated model, see for example Winkelmann [33]. Thus, moments, maximum likelihood estimators and a score test can be used to compare a zero-inflated quasi-binomial distribution with a zero-inflated binomial distribution, and a zero-inflated model with a homogeneous one. We illustrate this method using cross sectional data obtained by the Australian Health Survey (1977–78) and reported in Cameron and Trivedi [8, p. 67].

The rest of this paper is structured as follows. In order to make the paper as self-contained as possible, we first review the main properties of the quasi-binomial distribution, in Section 2. Section 3 then introduces the zero-inflated quasi-binomial model without covariates. Moment and maximum likelihood estimators of the proposed model are then considered and a score test is conducted to compare the two models: first, the zero-inflated quasi-binomial distribution against the zero-inflated binomial distribution, and then the zero-inflated model against the homogeneous model. Section 4 describes the inclusion of covariates in the above model. Some practical applications of the proposed

model are illustrated in Section 5, after which we summarise the main conclusions drawn.

2. Background

The quasi-binomial distribution (henceforth, QB distribution) was introduced by Consul [12] and later studied by Charalambides [10], Consul [12], Consul and Mittal [15] and Shenton [30]. This distribution depends on two parameters and includes as a particular case the classical binomial distribution when the second parameter takes the value zero. The QB distribution is useful for fitting over and under-dispersed data, an area in which it performs better than the traditional binomial model (see [12]).

Although various parameterisations of the QB distribution may be considered, in the present case we discuss the following, which has a useful probability mass function, given by

$$f(x;m,p,\psi)=\binom{m}{x}p(p+x\psi)^{x-1}(\bar{p}-x\psi)^{m-x},\quad x=0,1,\ldots,m, \tag{1}$$

where $x=0,1,\ldots$, $0\le p\le 1$, $-p/m<\psi\le(1-p)/m$ and $\bar{p}=1-p$. Hassan and Ahmed [25] extends the range of parameters of the quasi binomial distribution to allow $p+x\psi>1$ by taking

$$f(x;m,p,\psi)=\binom{m}{x}p(p+x\psi)^{x-1}(\bar{p}-x\psi)^{m-x},\quad x=0,1,\ldots,m-1,$$

and $f(x;m,p,\psi)=1-\sum_{j=0}^{m-1}f(x;m,p,\psi)$ for $x=m$. The parameter m is assumed to be fixed and known. Henceforth, this distribution is denoted as QB(p,ψ). Obviously, for $\psi=0$ the QB distribution is reduced to the classical binomial distribution with parameters m and p.

In the limit form, the QB distribution becomes the generalised Poisson distribution, which is a member of the Lagrangian family of distributions (see [10]).

Consul [12] provides the following simple interpretation of the parameters of the QB distribution. A random variable X following a QB distribution represents the number of successes in m trials such that the probability of the first success is p and in each successive trial it is $p+x\psi$. Therefore, the probability of success increases or decreases depending on whether $\psi>0$ or $\psi<0$. The mean and variance of the distribution are not given in simple expressions but they do not depend on any special function and can be computed in a straightforward way. They are given by

$$E_f(X;p,\psi)=mp\sum_{k=0}^{m-1}\psi^k(m-1)_{(k)}, \tag{2}$$

$$var_f(X;p,\psi)=mp\bar{p}-m(m-1)p^2\psi\times\sum_{k=0}^{m-2}\psi^k(m-2)_{(k)}((k+2)^2-km)$$

$$+\frac{1}{2}m(m-1)p\psi\sum_{k=0}^{m-2}(k+2)(k+3)\psi^k(m-2)_{(k)}$$
$$-m^2(m-1)^2p^2\psi^2\left[\sum_{k=0}^{m-2}\psi^k(m-2)_{(k)}\right]^2. \quad (3)$$

Here $(m-1)_{(k)} = (m-1)(m-2)\dots(m-k) = \Gamma(m)/\Gamma(m-k)$, where $\Gamma(y)$ represents the gamma function, $\Gamma(y) = \int_0^\infty t^{y-1}\,\mathrm{e}^{-t}\,\mathrm{d}t$, and $(m-1)_{(0)} = 1$.

3. Zero-inflated QB without covariates

Since the seminal paper by Cohen [11] many papers have addressed the subject of zero-inflated models. Starting from a discrete distribution $f(x)$, we can build a zero-inflated distribution in a simple form (see [11]) by writing,

$$g(x) = \begin{cases} w + (1-w)f(0), & x = 0, \\ 1-w)f(x), & x \neq 0, \end{cases} \quad (4)$$

where $f(x)$, $x = 0, 1, \dots$ is the parent distribution and $0 < w < 1$.

Now, replacing in (4) $f(x)$ by the QB distribution with probability mass function (1) we obtain the zero-inflated QB distribution, given by:

$$g(x; w, p, \psi) = \begin{cases} w + (1-w)\bar{p}^m, & x = 0, \\ (1-w)\binom{m}{x}p(p+x\psi)^{x-1}(\bar{p}-x\psi)^{m-x}, & x = 1, \dots, m. \end{cases} \quad (5)$$

It is convenient to write $w = \phi/(1+\phi)$ and the model (5) can then be represented as

$$g(x; \phi, p, \psi) = \begin{cases} \frac{1}{1+\phi}[\phi + \bar{p}^m], & x = 0, \\ \frac{1}{1+\phi}\binom{m}{x}p(p+x\psi)^{x-1}(\bar{p}-x\psi)^{m-x}, & x = 1, \dots, m, \end{cases} \quad (6)$$

where now it should be satisfied that $-(\phi+p)/m < \psi \leq (1-p+\phi)/m$ as it is easy to see.

Simple algebra provides the mean and variance of this distribution, given by

$$E_g(X; \phi, p, \psi) = \frac{1}{1+\phi}E_f(X; p, \psi), \quad (7)$$

$$\mathrm{var}_g(X; \phi, p, \psi) = \frac{1}{1+\phi}\,\mathrm{var}_f(X; p, \psi) + \frac{\phi}{(1+\phi)^2}E_f^2(X; p, \psi), \quad (8)$$

respectively, and where $E_f(X; p, \psi)$ and $\mathrm{var}_f(X; p, \psi)$ are given by (2) and (3), respectively.

3.1. Estimation

The distribution in (6) depends on three parameters. We assume that m is fixed and known, and therefore three equations are needed to estimate the parameters via the method of

moments. The mean-variance-and-zero-frequency is a simple procedure in this case, as shown by Alanko and Duffy [1]. If f_0 is the relative frequency of zeros in the sample we have

$$f_0 = \frac{1}{1+\phi}[\phi + \bar{p}^m].$$

From this equation we have $\phi = (f_0 - \bar{p}^m)/(1 - f_0)$, which can be extended to (7) and (8) to be solved numerically in order to obtain estimates of p and ψ.

In addition, based on a random sample $\tilde{X} = \{X_1, X_2, \ldots, X_n\}$ from (6), the log-likelihood is proportional to

$$\ell(\tilde{X}; \Theta) \propto n^* \log p - n \log(1+\phi) + n_0 \log[\phi + \bar{p}^m] + \sum_{x_i>0} [(x_i - 1)\log(p + x_i\psi) + (m - x_i)\log(\bar{p} - x_i\psi)], \tag{9}$$

where n_0 is the number of zeros in the sample, $n^* = n - n_0$ and $\Theta = (\phi, p, \psi)$ is the vector of parameters which has to be estimated.

From (9) it is straightforward to derive the following normal equations:

$$\frac{\partial \ell(\tilde{X}; \Theta)}{\partial \phi} = \frac{-n}{1+\phi} + \frac{n_0}{\phi + \bar{p}^m} = 0, \tag{10}$$

$$\frac{\partial \ell(\tilde{X}; \Theta)}{\partial p} = \frac{n^*}{p} - \frac{n_0 m \bar{p}^{m-1}}{\phi + \bar{p}^m} + \sum_{x_i>0} \frac{x_i - 1}{p + x_i\psi} - \sum_{x_i>0} \frac{m - x_i}{\bar{p} - x_i\psi} = 0, \tag{11}$$

$$\frac{\partial \ell(\tilde{X}; \Theta)}{\partial \psi} = \sum_{x_i>0} \frac{(x_i - 1)x_i}{p + x_i\psi} - \sum_{x_i>0} \frac{(m - x_i)x_i}{\bar{p} - x_i\psi} = 0. \tag{12}$$

From (10) we have $\phi = [n_0 - n\bar{p}^m]/(n - n_0)$ which can be extended to equations (11) and (12) to obtain a system of equations which depends on p and ψ and can be solved numerically.

The second order partial derivatives are as follows,

$$\frac{\partial^2 \ell(\tilde{X}; \Theta)}{\partial \phi^2} = \frac{n}{(1+\phi)^2} - \frac{n_0}{[\phi + \bar{p}^m]^2},$$

$$\frac{\partial^2 \ell(\tilde{X}; \Theta)}{\partial \phi \partial p} = \frac{n_0 m (1-p)^{m-1}}{[\phi + \bar{p}^m]^2}, \quad \frac{\partial^2 \ell(\tilde{X}; \Theta)}{\partial \phi \partial \psi} = 0,$$

$$\frac{\partial^2 \ell(\tilde{X}; \Theta)}{\partial p^2} = \frac{n_0 m \bar{p}^{m-2}[(m-1)\phi - \bar{p}^m]}{[\phi + \bar{p}^m]^2} - \frac{n^*}{p^2} - \sum_{x_i>0} \frac{x_i - 1}{(p + x_i\psi)^2} - \sum_{x_i>0} \frac{m - x_i}{(\bar{p} - x_i\psi)^2},$$

$$\frac{\partial^2 \ell(\tilde{X}; \Theta)}{\partial p \partial \psi} = -\sum_{x_i>0} \frac{(x_i - 1)x_i}{(p + x_i\psi)^2} - \sum_{x_i>0} \frac{(m - x_i)x_i}{(\bar{p} - x_i\psi)^2},$$

$$\frac{\partial^2 \ell(\tilde{X}; \Theta)}{\partial \psi^2} = -\sum_{x_i>0} \frac{(x_i - 1)x_i^2}{(p + x_i\psi)^2} - \sum_{x_i>0} \frac{(m - x_i)x_i^2}{(\bar{p} - x_i\psi)^2}.$$

To obtain the Fisher's information matrix we need the expectations given in Appendix A and which are described in Consul [12]. These elements provide the elements of the matrix in a closed form. Then, the elements of Fisher's information matrix, $I(\Theta) = (I_{i,j})_{i,j=1,2,3}$, are given as follows:

$$
\begin{aligned}
I_{11} &= E\left(-\frac{\partial^2 \ell(\tilde{X};\Theta)}{\partial \phi^2}\right) = -\frac{n}{(1+\phi)^2} + \frac{n_0}{[\phi+\bar{p}^m]^2},\\
I_{12} &= E\left(-\frac{\partial^2 \ell(\tilde{X};\Theta)}{\partial \phi \partial p}\right) = I_{21} = E\left(-\frac{\partial^2 \ell(\tilde{X};\Theta)}{\partial p \partial \phi}\right)\\
&= -\frac{n_0 m \bar{p}^{m-1}}{[\phi+\bar{p}^m]^2},\\
I_{13} &= E\left(-\frac{\partial^2 \ell(\tilde{X};\Theta)}{\partial \phi \partial \psi}\right) = I_{31} = E\left(-\frac{\partial^2 \ell}{\partial \psi \partial \phi}\right) = 0,\\
I_{22} &= E\left(-\frac{\partial^2 \ell(\tilde{X};\Theta)}{\partial p^2}\right) = \frac{n_0 m \bar{p}^{m-2}[(m-1)\phi - \bar{p}^m]}{[\phi+\bar{p}^m]^2} + \frac{n^* m(1+(1-m)\psi)}{\bar{p}+(1-m)\psi}\\
&\quad + \frac{n^* m[2\psi + p(1-p(1-m)+(1-m)\psi)}{p(p+2\psi)},\\
I_{23} &= E\left(-\frac{\partial^2 \ell(\tilde{X};\Theta)}{\partial p \partial \psi}\right) = I_{32} = \frac{n^* p m_{(2)}[1-(m-3)\psi]}{(p+2\psi)(\bar{p}+(1-m)\psi)},\\
I_{33} &= E\left(-\frac{\partial^2 \ell(\tilde{X};\Theta)}{\partial \psi^2}\right) = \frac{n^* m_{(2)} p[2+p(m-3)]}{(p+2\psi)(\bar{p}+(1-m)\psi)}.
\end{aligned}
$$

3.2. Simulation experiment

We now perform some simulation experiments via bootstrapping method [19,32], among others, to examine the behaviour of the maximum likelihood estimators and the correlation between these estimators. In this study we have used the `Mathematica software package` to generate directly random variates from the pdf (1) by using the aforementioned stochastic representation.

The sample sizes used in this simulation analysis to compute the estimates were $n = 500$ and $n = 750$. The average estimates and square root of the mean squared errors were calculated based on 1000 replications. Additional replications seem to be unnecessary, as computational time would be prohibitive, even though fewer replications might reduce the statistical accuracy obtained. Clearly, as the sample size increases, the biases and the mean squared errors tend to decrease, which seems to verify the consistency properties of the maximum likelihood estimators. These results together with the correlation ρ, between the parameters for the distribution, the bias and average mean square error (MSE) of the simulated estimates are shown in Table 1. It can be seen that the 95% percentile bootstrap confidence interval (CI) of the mean of the estimated parameter covers the true value of the parameters.

Table 1. Average estimates (first row), the square root of the mean squared errors (second row in parentheses) for the distribution proposed based on 500 replications and correlation, ρ, between the parameters.

	$m = 10$		$m = 15$	
n	$p = 0.10$	$\psi = 0.05$	$p = 0.50$	$\psi = 0.02$
500	0.092961	0.047638	0.499951	0.018542
	(0.0052)	(0.0029)	(0.0171)	(0.0016)
Average bias	−0.007038	−0.002361	−0.000049	−0.001457
Average MSE	0.000077	0.000014	0.000293	4.95315E−6
95% CI	(0.0891, 0.1032)	(0.0455, 0.0534)	(0.4874, 0.5340)	(0.0174, 0.0216)
	$\rho = -0.4128$		$\rho = -0.9521$	
750	0.099907	0.048953	0.494508	0.018853
	(0.0043)	(0.0024)	(0.0136)	(0.0013)
Average bias	−0.000092	−0.001046	−0.005491	−0.001146
Average MSE	0.000019	7.2193E−6	0.000215	3.0212E−6
95% CI	(0.0971, 0.1084)	(0.0471, 0.0538)	(0.4856, 0.5216)	(0.0180, 0.0212)
	$\rho = -0.4262$		$\rho = -0.9494$	

3.3. Introducing two score tests

In this section, we present two score tests, used to compare the zero-inflated QB distribution with the zero-inflated binomial distribution, and the zero-inflated QB model with the homogeneous QB model. The reader is referred to Cox and Hinkley [16], Broek [3] and Gupta et al. [22] for a discussion of the score test.

As is well known, the ZIQB distribution model in (6) reduces to the ZIB model when the parameter $\psi = 0$. To test the adequacy of the ZIQB distribution model with respect to the ZIB model, we consider the hypothesis $H_0 : \psi = 0$ versus $H_1 : \psi \neq 0$. The inclusion of the parameter ψ in the model is justified if H_0 is rejected.

The score vector obtained from (10)–(12) is $\mathbf{U}_1 = (\Xi_1, \Xi_2, \Xi_3)$, where

$$\Xi_1 = \frac{-n}{1+\widehat{\phi}} + \frac{n_0}{\widehat{\phi}+\widehat{\overline{p}}^m},$$

$$\Xi_2 = -\frac{n_0 m \widehat{\overline{p}}^{m-1}}{\widehat{\phi}+\widehat{\overline{p}}^m} + \frac{1}{\widehat{\overline{p}}\widehat{\overline{p}}}(n\bar{x} - n^* m\widehat{p}),$$

$$\Xi_3 = \frac{n[s^2 + \bar{x}^2 - \bar{x}(m\widehat{p} + \widehat{\overline{p}})]}{\widehat{\overline{p}}\widehat{\overline{p}}},$$

and where $\bar{x}$ and s^2 represent the sample mean and the variance, respectively and where $\hat{p}$ and $\hat{\phi}$ are the maximum likelihood estimates of p and ϕ, respectively. From $J \equiv J_{i,j}(\hat{\phi}, \hat{p}, 0) = I^{-1}_{i,j}(\hat{\phi}, \hat{p}, 0)$, it is well known that the statistics $W = \mathbf{U}_1 J \mathbf{U}_1'$, under the null hypothesis, have an asymptotic chi-squared distribution with one degree of freedom.

To test the adequacy of the homogeneous model with respect to the zero-inflated model, we consider the hypothesis $H_0 : \phi = 0$ versus $H_1 : \phi \neq 0$. In this case, the score vector is $\mathbf{U}_2 = (\Upsilon_1, \Upsilon_2, \Upsilon_3)$, where

$$\Upsilon_1 = \frac{-n\widehat{\overline{p}}^m + n_0}{\widehat{\overline{p}}^m},$$

$$\Upsilon_2 = \frac{-n_0 m\widehat{p} + n^*\widehat{\bar{p}}}{\widehat{p}\widehat{\bar{p}}} + \sum_{x_i>0}\left(\frac{x_i - 1}{\widehat{p} + x_i\widehat{\psi}} - \frac{m - x_i}{\widehat{\bar{p}} - x_i\widehat{\psi}}\right),$$

$$\Upsilon_3 = \sum_{x_i>0}\frac{(x_i - 1)x_i}{\widehat{p} + x_i\widehat{\psi}} - \sum_{x_i>0}\frac{(m - x_i)x_i}{\widehat{\bar{p}} - x_i\widehat{\psi}}.$$

Again, the statistic $T = \mathbf{U}_1 J \mathbf{U}_1'$, under the null hypothesis, has an asymptotic chi-squared distribution with one degree of freedom. Here, $J \equiv J_{i,j}(0, \hat{p}, \hat{\psi}) = I^{-1}_{i,j}(0, \hat{p}, \hat{\psi})$, where $\hat{p}$ and $\hat{\psi}$ are the maximum likelihood estimates of p and ψ, respectively.

Because of the complex form of the Fisher information matrix, both statistics require more effort to present them in a simple formula.

4. The inclusion of covariates

A new reparameterisation of the QB distribution in (1) can be obtained from (2) by replacing p by the expression

$$p(\theta, \psi) = \frac{\theta}{m\sum_{k=0}^{m-1}\psi^k(m-1)_{(k)}} = \frac{\theta}{m}\frac{\psi^{1-m}\,\mathrm{e}^{-1/\psi}}{\Gamma(m, 1/\psi)}, \tag{13}$$

from which we obtain the probability density function

$$g(x; \theta, \psi) = p(\theta, \psi)\left(p(\theta, \psi) + x\psi\right)^{x-1}\left(\bar{p}(\theta, \psi) - x\psi\right)^{m-x}, \quad x = 0, 1, \ldots, m, \tag{14}$$

where $\bar{p}(\theta_i, \psi) = 1 - p(\theta_i, \psi)$, and where the parameter θ, i.e. $E_g(X; \theta, \psi) = \theta$, is now the mean of the distribution in (14). The operation is carried out in this way because the practitioner usually wishes to include covariates in the model, and this is the most convenient way to do so.

Under this modification, the model (6) can be written as:

$$g(x_i; \phi, \theta_i, \psi) = \begin{cases} \dfrac{1}{1+\phi}\{\phi + [1 - p(\theta_i, \psi)]^m\}, & x_i = 0, \\ \dfrac{1}{1+\phi}\binom{m}{x_i}p(\theta_i, \psi)[p(\theta_i, \psi) + x_i\psi]^{x_i-1} & \\ \quad \times[\bar{p}(\theta_i, \psi) - x_i\psi]^{m-x_i}, & x_i = 1, \ldots, m,\ i = 1, \ldots, n. \end{cases} \tag{15}$$

In this case, ψ is interpreted as a precision parameter in the sense that, for fixed θ, the larger the value of ψ, the larger the variance of X. This can be seen by observing the numerical values in Table 2, which shows the values of the variance for different values of the parameters m, θ and ψ.

The most common specification for the mean parameter θ is exponential, which ensures the nonnegativity of θ. That is,

$$\theta_i = \frac{m\exp(\boldsymbol{y}^\top\boldsymbol{\beta})}{1 + \exp(\boldsymbol{y}^\top\boldsymbol{\beta})}, \quad i = 1, \ldots, n,$$

where $\boldsymbol{y}_i^\top = (y_{i1}, y_{i2}, \ldots, y_{iq})$ are observations of q known covariates (independent variables or regressors) and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top \in \mathbb{R}^q$ is a q-vector of unknown regression

Table 2. Values of the variance of the distribution for two values of θ considered as fixed.

m	θ	ψ	var(X)
10	0.5	0.0075	0.544
10	0.5	0.075	1.992
10	0.5	0.75	21.028
10	0.5	1.00	24.475
5	2	0.0025	1.224
5	2	0.025	1.463
5	2	0.25	5.721
5	2	1.00	17.114

coefficients. Then, we have the conventional log-linear model such that $E(X_i) = \theta_i \in [0, m]$ and so it, too, is bounded.

The log-likelihood of the ZIQB distribution model with covariates, obtained from (15), is proportional to

$$\begin{aligned}\ell(\phi, \boldsymbol{\beta}, \psi) \propto & -n\log(1+\phi) + \sum_{x_i=0} \log\{\phi + [\bar{p}(\theta_i, \psi)]^m\} \\ & + \sum_{x_i>0} \log p(\theta_i, \psi) + \sum_{x_i>0} (x_i - 1)\log[p(\theta_i, \psi) + x_i\psi] \\ & + \sum_{x_i>0} (m - x_i)\log[\bar{p}(\theta_i, \psi) - x_i\psi].\end{aligned}$$

The normal equations are given by,

$$\frac{\partial \ell(\phi, \boldsymbol{\beta}, \psi)}{\partial \phi} = -\frac{n}{1+\phi} + \sum_{x_i=0} \frac{1}{\phi + [\bar{p}(\theta_i, \psi]^m} = 0, \tag{16}$$

$$\begin{aligned}\frac{\partial \ell(\phi, \boldsymbol{\beta}, \psi)}{\partial \beta_s} = & -m \sum_{x_i=0} \frac{\partial p(\theta_i, \psi)}{\partial \beta_s} \frac{[\bar{p}(\theta_i, \psi)]^{m-1}}{\phi + [\bar{p}(\theta_i, \psi)]^m} \\ & + \sum_{x_i>0} \frac{\partial p(\theta_i, \psi)}{\partial \beta_s} \left[\frac{1}{p(\theta_i, \psi)} + \frac{x_i - 1}{p(\theta_i, \psi) + x_i\psi}\right] \\ & - \sum_{x_i>0} \frac{\partial p(\theta_i, \psi)}{\partial \beta_s} \frac{m - x_i}{\bar{p}(\theta_i, \psi) - x_i\psi} = 0, \quad s = 1, \ldots, q,\end{aligned} \tag{17}$$

$$\begin{aligned}\frac{\partial \ell(\phi, \boldsymbol{\beta}, \psi)}{\partial \psi} = & -m \sum_{x_i=0} \frac{\partial p(\theta_i, \psi)}{\partial \psi} \frac{[\bar{p}(\theta_i, \psi)]^{m-1}}{\phi + [\bar{p}(\theta_i, \psi)]^m} \\ & + \sum_{x_i>0} \frac{\partial p(\theta_i, \psi)}{\partial \psi} \left[\frac{1}{p(\theta_i, \psi)} + \frac{x_i - 1}{p(\theta_i, \psi) + x_i\psi}\right] \\ & - \sum_{x_i>0} \left[x_i + \frac{\partial p(\theta_i, \psi)}{\partial \psi}\right] \frac{m - x_i}{\bar{p}(\theta_i, \psi) - x_i\psi} = 0,\end{aligned} \tag{18}$$

where,

$$\frac{\partial p(\theta_i, \psi)}{\partial \beta_s} = (m - \theta_i)\theta_i y_s p(\theta_i, \psi),$$
$$\frac{\partial p(\theta_i, \psi)}{\partial \psi} = \frac{p\{mp + \theta[(m-1)\psi - 1]\}}{\theta \psi^2}.$$

Maximum likelihood estimates of ϕ, ψ and $\beta_s, s = 1, \ldots q$, can be obtained via the Newton-Raphson method for both non-inflated and inflated models, with and without covariates. In this case, we estimate the parameters by the maximum likelihood method, using the standard procedure in the Mathematica (see [29]), RATS (see [4]) and R statistical packages. A simple program developed with R software is provided, for the case of the ZIQB model, in Appendix B. Some details about the estimation procedure are also provided in the following section.

4.1. Interpretation of the regressors

The marginal effect reflects the variation of the conditional mean of X due to a one-unit change in the sth covariate, and is calculated as

$$\frac{\partial \theta_i}{\partial y_s} = \frac{1}{m}(m - \theta_i)\beta_s,$$

for $i = 1, \ldots, n$ and $s = 1, \ldots, q$. Thus, the marginal effect indicates that a one-unit change in the sth regressor increases or decreases the expectation of the dependent variable, depending on the sign, positive or negative, of the regressor for each mean. For indicator variables such as y_s, which takes only the value 0 or 1, the marginal effect in terms of the odds-ratio is approximately $\exp(\beta_j)$. Therefore, the conditional mean is $\exp(\beta_j)$ times larger if the indicator variable is one rather than zero.

5. Applications

In this section, we consider two numerical applications to show how the proposed models perform in comparison with classical models previously described in the literature. In our comparisons, the following distributions are considered: a Poisson (P) distribution with parameter $\theta > 0$; a negative binomial (NB) distribution with parameters $\alpha > 0$ and mean $\theta > 0$; a generalised Poisson (GP) distribution with parameters $\alpha > 0$ and mean $\theta > 0$ and of course the QB distribution with parameters ψ and θ. Among the various parameterisations of the generalised Poisson distribution, we used the one described in [14].

The values of T and W for the data considered are very strongly significative against the homogeneous model and against the inflated binomial model. Accordingly, these models were not considered in our comparison.

The different models considered were analysed using the BFGS algorithm (see [5,6,21]), with RATS (see [4]), Wolfram Mathematica and R software (see [29]), for both the inflated and the non-inflated models. In all of the models considered, the convergence of the algorithm is extremely fast. In particular, for the ZIQB model (see Appendix C), the algorithm converged in fewer than 30 iterations.

5.1. Application to the Australian health survey data

5.1.1. Description of the data

This study is based on data obtained from the 1977–78 Australian Health Survey, a well-known data set previously studied by Cameron and Trivedi [8]; see also Cameron and Trivedi [7]. The data can be downloaded from the web page http://cameron.econ.ucdavis.edu/racd/racddata.html.

Details about this data source can also be consulted in the 'Ecdat' R (data(DoctorAUS)) package. The data set consists of 5190 elements (85.81% of which are zeros) with fifteen variates. The variable ACTDAYS (the dependent variable) represents the number of days of reduced activity in the past two weeks due to illness or injury. Therefore this variable has support in the interval [0, 14] and is bounded. In our study, CHCOND (chronic condition) is not considered, and INSURANCE (medlevy : medibank levy, levyplus : private health insurance, freepoor: government insurance due to low income, freerepa : government insurance due to old age disability or veteran status) is converted into three dichotomous variables, FREEPOR, FREEREPA AND LEVYPLUS. Therefore, MEDLEVY is the reference variable. The main target of this study is to explain the ACTDAYS variate in terms of the insurance covariate to assess the nexus between the insurance level and the use of medical license permit (a better insurance contract promotes a higher number of days with reduced activity). Descriptive statistics on the variables in this dataset are given in Table 3.2 of Cameron and Trivedi [7, p.68].

5.1.2. Results

Table 3 shows the characteristics of the fit obtained in this example. In this respect, the maximum likelihood method was used under the homogeneous model (without covariates) both with and without the inflation of zeros. As can be seen, all the parameters are statistically significant and the QB distribution provides the better fit in each of the cases considered.

Tables 4 and 5 show the estimation in the case of the non-homogeneous model (i.e. considering the inclusion of covariates) with and without inflation. Again, in view of the maximum value of the logarithm of the likelihood function, the QB model is clearly superior.

5.1.3. Interpretation

In general, most of the variables are highly significant, although there are slight differences between the models, and also when a distinction is made between the inflated and non-inflated models. The high percentage of zeros in the sample is reflected in the fact that the three covariables corresponding to the type of insurance that covers the patient are not significant, and neither is the number of days of hospital stay. Finally, we provide an interpretation of the estimated parameters in both applications according to the comments provided in Subsection 4.1. For the first application, keeping fixed the rest of covariates, the mean for ACTDAYS is increased 1.028 times for people with government insurance due to low income and private health insurance, as compared to people with medibank levy insurance. Similarly, the mean for ACTDAYS decreases 0.846 times for people with government insurance due to old age disability or veteran status, when compared to persons with medibank levy insurance. Also, women increase 1.036 times the mean for ACTDAYS

Table 3. Fitted homogeneous models and homogeneous inflated models without covariates for Australian health survey data. The standard errors of the estimate parameters are shown in brackets.

		Fitted							
Counts	Observed	P	ZIP	NB	ZINB	GP	ZIGP	QB	ZIQB
0	4454	2194.02	4454.00	4445.93	4454.00	4434.26	4454.00	4445.45	4454.00
1	177	1889.05	10.39	229.85	152.17	308.14	145.57	248.30	189.56
2	108	813.23	31.52	113.97	106.88	121.93	113.37	101.75	94.96
3	74	233.39	63.71	73.38	81.38	68.19	87.41	59.91	60.63
4	45	50.24	96.57	52.68	64.37	44.49	68.10	41.44	43.81
5	40	8.65	117.10	40.16	52.05	31.68	53.76	31.46	34.17
6	17	1.24	118.33	31.81	42.71	23.87	42.99	25.42	28.10
7	38	0.15	102.49	25.86	35.40	18.71	34.79	21.51	24.05
8	17	0.01	77.67	21.44	29.57	15.10	28.45	18.88	21.27
9	7	0.00	52.33	18.03	24.84	12.47	23.47	17.11	19.35
10	12	0.00	31.72	15.35	20.97	10.48	19.52	15.96	18.03
11	2	0.00	17.48	13.19	17.77	8.939	16.35	15.28	17.11
12	6	0.00	8.83	11.42	15.10	7.714	13.77	14.75	15.92
13	5	0.00	4.12	9.95	12.87	6.725	11.67	9.44	4.87
14	188	0.00	1.78	8.72	11.00	5.913	9.994	123.27	164.00
Total	5190.00	5190.00	5188.86	5111.82	5125.45	5111.68	5122.95	5190.00	5190.00
ϕ			6.035		4.061		4.752		2.389
			(0.236)		(0.317)		(0.241)		(0.403)
$\widehat{\theta}$		0.861	6.063	0.862	4.361	0.862	4.958		
		(0.012)	(0.090)	(0.051)	(0.304)	(0.070)	(0.264)		
$\widehat{\alpha}$				0.055	0.597	5.194	0.389		
				(0.002)	(0.085)	(0.210)	(0.034)		
$\widehat{\psi}$								0.076	0.073
								(0.000)	(0.000)
$\widehat{p}$								0.011	0.0457
								(0.000)	(0.007)
$\ell_{\max}$		−11570.568	−4950.69	−4141.242	−4110.34	−4224.884	−4126.25	−3678.938	−3667.34

when compared to men. The mean value for ACTDAYS is increased 2.368 times for each consultation with a doctor or specialist in the past 2 weeks and is increased 1.542 times for each admission to a hospital, psychiatric hospital, nursing or convalescent home in the past 12 months.

5.1.4. Checking the models

As with ordinary linear regression, metrics such as the mean absolute error (MAE) and the root of mean squared error (RMSE) are commonly used to measure accuracy for continuous and discrete variables. Additionally, residuals can be used in order to study the fit of the model to the data and in particular to detect outliers and to test the variance assumption. These metrics are defined by

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i|,$$

$$\text{RMSE} = \left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2\right)^{1/2},$$

$$\text{Raw residuals} = \theta_i - \widehat{\theta}_i, \quad i = 1, \ldots, n.$$

The MAE and RMSE values for all the models considered are shown in Tables 4 and 5. In most of the cases the QB model has the lowest values of these two metrics. Figure 1 shows

Table 4. Coefficient estimates and *p*-values for the homogeneous models with covariates and Australian health survey data.

	P		NB		GP		QB	
Variable	Estimate	Pr $>\|t\|$	Variable	Pr $>\|t\|$	Variable	Pr $>\|t\|$	Variable	Pr $>\|t\|$
SEX	0.059	0.032	−0.021	0.828	−0.055	0.675	0.035	0.678
AGE	0.346	0.000	0.080	0.793	−0.717	0.046	−1.610	0.000
INCOME	0.305	0.000	0.178	0.197	−0.045	0.785	−0.054	0.666
ILLNESS	0.159	0.000	0.376	0.000	0.824	0.000	0.267	0.000
HSCORE	0.126	0.000	0.159	0.000	0.159	0.000	0.155	0.000
DOCTORCO	0.293	0.000	0.790	0.000	1.614	0.000	0.444	0.000
NONDOCCO	0.101	0.000	0.281	0.000	1.102	0.000	0.309	0.000
HOSPADMI	0.020	0.249	0.330	0.000	0.272	0.040	0.120	0.009
HOSPDAYS	0.016	0.000	0.027	0.004	0.121	0.000	0.012	0.017
MEDECINE	0.044	0.000	0.076	0.006	0.104	0.000	0.131	0.000
PRESCRIB	0.044	0.000	0.073	0.032	0.107	0.018	0.148	0.000
NONPRESC	0.084	0.000	0.147	0.035	0.173	0.038	0.168	0.000
FREEPOR	−0.126	0.000	−0.119	0.210	−0.089	0.211	−0.056	0.367
FREEREPA	−0.125	0.005	0.040	0.813	0.844	0.000	−0.272	0.082
LEVYPLUS	−0.126	0.000	−0.119	0.210	−0.089	0.211	−0.056	0.367
CONSTANT	−1.376	0.000	−2.319	0.000	−2.856	0.000	−3.242	0.000
$\widehat{\alpha}$			0.094	0.000	2.976	0.000		
$\widehat{\psi}$							0.085	0.000
$\ell_{\max}$	−8709.189		−3869.711		−3958.239		−3013.919	
MAE	1.225		41.36		>100		1.356	
RMSE	2.937		>100		>100		2.603	

Table 5. Coefficient estimates and *p*-values for the zero inflated models with covariates and Australian health survey data.

	P		NB		GP		QB	
Variable	Estimate	Pr $>\|t\|$	Variable	Pr $>\|t\|$	Variable	Pr $>\|t\|$	Variable	Pr $>\|t\|$
SEX	−0.099	0.002	−0.027	0.794	−0.020	0.868	−0.089	0.385
AGE	1.286	0.000	−0.101	0.743	−0.210	0.589	−1.424	0.000
INCOME	0.097	0.043	0.180	0.203	−0.031	0.844	0.114	0.411
ILLNESS	−0.027	0.020	0.394	0.000	0.633	0.000	0.314	0.000
HSCORE	0.055	0.000	0.174	0.000	0.155	0.000	0.167	0.000
DOCTORCO	0.084	0.000	1.091	0.000	1.309	0.000	0.862	0.000
NONDOCCO	0.061	0.000	0.319	0.000	0.966	0.000	0.412	0.000
HOSPADMI	0.082	0.000	0.409	0.000	0.288	0.020	0.433	0.000
HOSPDAYS	0.005	0.000	0.023	0.005	0.112	0.000	0.026	0.043
MEDECINE	0.006	0.153	0.082	0.000	0.083	0.001	0.101	0.000
PRESCRIB	0.006	0.170	0.080	0.001	0.083	0.032	0.146	0.000
NONPRESC	−5E−4	0.973	0.160	0.005	0.139	0.039	0.228	0.000
FREEPOR	−0.101	0.000	−0.112	0.072	−0.107	0.093	0.028	0.625
FREEREPA	0.011	0.822	0.026	0.875	0.791	0.000	−0.167	0.356
LEVYPLUS	−0.101	0.000	−0.112	0.072	−0.107	0.093	0.028	0.625
CONSTANT	0.995	0.000	−8.131	0.000	−1.665	0.000	−3.534	0.000
$\widehat{\psi}$							0.081	0.000
$\widehat{\alpha}$			2.6E−3	0.193	0.759	0.000		
$\widehat{\phi}$	5.961	0.000	0.143	0.000	1.768	0.000	0.345	0.000
$\ell_{\max}$	−4417.017		−3815.712		−3929.91		−2986.297	
MAE	4.579		3.825		>100		1.511	
RMSE	4.957		129.47		>100		2.796	

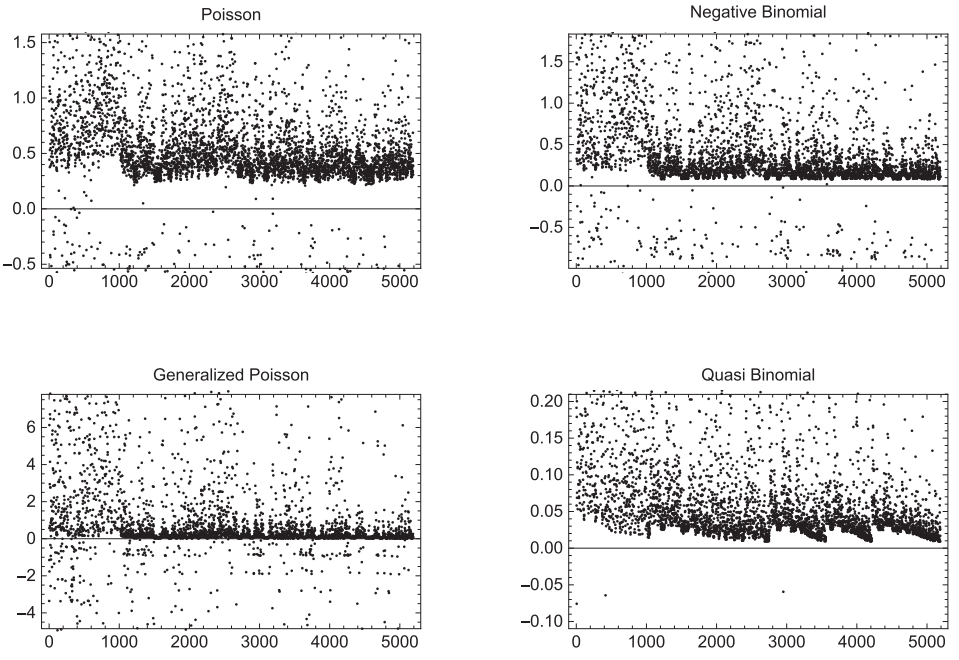

Figure 1. Raw residuals for the models considered. Australian health survey data.

the raw residuals for the models considered, revealing a greater dispersion of the Poisson, negative binomial and generalised Poisson distributions than of the QB distribution. In the latter case, the model always underestimates, with values not exceeding approximately 0.17.

As well as determining the maximum value of the logarithm of the likelihood function, the MAE, the RMSE and the raw residuals, we also performed the Vuong test to distinguish between each of the models considered, comparing the estimates of the quasi binomial (QB) with respect to the Poisson, negative binomial (NB) and generalised Poisson (GP) models, all of which are non-nested with the QB model. In this regard, we tested the null hypothesis that the two models are equally close to the actual model, against the alternative that one of them is closer (see [[18, p.43], [31]]). The z-statistic is

$$Z = \frac{1}{\omega\sqrt{n}}(\ell(\widehat{\theta_1}) - \ell(\widehat{\theta_2})),$$

where $\widehat{\theta_1}$ and $\widehat{\theta_2}$ are vectors of the parameters considered and

$$\omega^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\log\left(\frac{f(x_i|\widehat{\theta_1})}{g(x_i|\widehat{\theta_2})}\right)\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f(\widehat{\theta_1})}{g(\widehat{\theta_2})}\right)\right]^2$$

where f and g represent the QB and the Poisson (negative binomial) distributions, respectively.

Due to the asymptotically normal behaviour of the Z statistic under the null hypothesis, this hypothesis is rejected in favour of the alternative, i.e. that model $f(\cdot)$ occurs with a

Table 6. Vuong tests comparing the QB model with the Poisson, NB and GP models for Australian health survey data. The *p*-value is shown below.

	Homogeneous model		Inflated model	
	Without covariates	With covariates	Without covariates	With covariates
QB vs. Poisson	23.53	18.81	17.55	14.49
	< 0.001	< 0.001	< 0.001	< 0.001
QB vs. NB	12.24	10.93	12.02	4.07
	< 0.001	< 0.001	< 0.001	< 0.001
QB vs. GP	12.72	10.78	11.97	11.26
	< 0.001	< 0.001	< 0.001	< 0.001

significance level α, when $Z > z_{1-\alpha}$, where $z_{1-\alpha}$ is the $(1-\alpha)$ quantile of the standard normal distribution.

From the values obtained in this test (see Table 6) and the corresponding *p*-values (in brackets), we can discriminate between the three competing models according to the data available, concluding that the QB model achieves the best fit to the model.

5.2. Application to a baseball data

5.2.1. Description of the data

This another study is based on data which includes information for 3402 individual pitches thrown by Los Angeles Dodger baseball pitcher Clayton Kershaw during the 2013 regular season when he won the Cy Young award as the best pitcher in the National League. The data can be downloaded from the web page http://gd2.mlb.com/components/game/mlb/.

Details about this data source can also be consulted in the R Package 'Stat2Data' package. The variables considered here are the followings:

- BallCount: Number of balls before the pitch. This is now the dependent variable, which is bounded between 0 and 3.
- BatterNumber: Number of batters faced so far that game.
- Swing: Did the batter swing at the pitch? (No or Yes).
- StartSpeed: Speed leaving the pitcher's hand (in miles per hour, mph).
- EndSpeed: Speed crossing home plate (in mph).
- Zone: 1–9 in theoretical strike zone (upper left to lower right), 11–14 are out of strike zone 82 KeyWestWater.
- Nasty: A measure on a 0-100 scale of difficulty of the pitch to hit (100 is most difficult).
- Count Ball strike count (0-0, 0-1, 0-2, 1-1, 1-2, 2-1, 2-2, 3-1, or 3-2).
- StrikeCount: Number of strikes before the pitch (0, 1, or 2).
- Inning: Inning of the game.
- Outs: Number of outs when the pitch is thrown.
- BatterHand: Batter's stance (L = left or R = right).

For this problem, the main issue is to explain the BallCount variable in terms of the rest of available information to assess which variables increases the number of balls before the pitch.

5.2.2. Results

Tables 7–9 show the characteristics of the fit obtained in this example for the cases of not including covariates and including covariates. Again most of the parameters result statistically significant and the QB distribution provides the better fit in each of the cases considered.

5.2.3. Checking the models

Again, MAE and RMSE metrics for all the models considered are shown in Tables 8 and 9 providing the lowest values of them fo the models based on the QB distribution.

Table 7. Fitted homogeneous models and homogeneous inflated models without covariates for baseball data. The standard errors of the estimate parameters are shown in brackets.

		Fitted							
Counts	Observed	P	ZIP	NB	ZINB	GP	ZIGP	QB	ZIQB
0	1569	1439.06	1569.00	1497.15	1569.00	1493.30	1569.00	1575.74	1569.00
1	1013	1238.13	1052.92	1173.18	1053.21	1177.80	966.62	986.86	1013.00
2	546	562.63	538.26	512.00	538.20	513.00	659.49	574.74	543.25
3	274	152.75	183.44	164.191	183.35	163.38	186.65	264.02	270.04
Total	3402.00	3362.57	3343.62	3346.53	3343.67	3347.46	3381.77	3402.00	3402.00
ϕ			0.188		0.188		0.459		−0.187
			(0.027)		(0.027)		(0.038)		(0.129)
$\widehat{\theta}$		0.860	1.022	0.860	1.022	0.860	1.255		
		(0.016)	(0.030)	(0.016)	(0.016)	(0.019)	(0.034)		
$\widehat{\alpha}$				8.782	59210.49	0.052	−0.148		
				(2.796)	(132.40)	(0.017)	(0.012)		
$\widehat{\psi}$								0.120	0.145
								(0.006)	(0.017)
$\widehat{p}$								0.226	0.174
								(0.005)	(0.034)
$\ell_{\max}$		−4236.579	−4209.179	−4230.846	−4209.181	−4231.224	−4180.133	−4131.846	−4130.595

Table 8. Coefficient estimates and *p*-values for the homogeneous models with covariates for baseball data.

	P		NB		GP		QB	
Variable	Estimate	Pr > |t|	Variable	Pr > |t|	Variable	Pr > |t|	Variable	Pr > |t|
Batter number	0.006	0.123	0.007	0.088	0.006	0.122	0.011	0.084
Swing	0.004	0.910	0.006	0.880	0.001	0.972	0.005	0.923
log(Start speed)	0.148	0.253	1.386	0.000	0.294	0.000	0.067	0.717
log(End speed)	0.222	0.070	−1.051	0.000	0.050	0.000	0.482	0.002
log(zone)	−0.013	0.577	−0.012	0.633	−0.016	0.395	−0.023	0.513
log(Nasty)	−0.064	0.174	−0.066	0.000	−0.059	0.000	−0.102	0.111
Strike count	0.872	0.000	0.872	0.000	0.833	0.000	1.300	0.000
Inning	−0.063	0.075	−0.070	0.053	−0.062	0.069	−0.104	0.047
Outs	0.008	0.648	0.006	0.719	0.007	0.657	0.012	0.624
Batter hand	0.063	0.140	0.065	0.167	0.062	0.121	0.113	0.059
CONSTANT	−2.647	0.015	−2.583	0.000	−2.497	0.000	−4.467	0.002
$\widehat{\alpha}$			10997.827	0.000	−0.089	0.000		
$\widehat{\psi}$							0.072	0.000
$\ell_{\max}$	−3691.117		−3691.197		−3667.994		−3531.917	
MAE	0.671		0.671		0.673		0.654	
RMSE	0.852		0.851		0.849		0.848	

Table 9. Coefficient estimates and *p*-values for the zero inflated models with covariates for baseball data.

	P		NB		GP		QB	
Variable	Estimate	Pr > |t|	Variable	Pr > |t|	Variable	Pr > |t|	Variable	Pr > |t|
Batter number	0.006	0.117	0.007	0.116	0.006	0.079	0.011	0.076
Swing	−2.5E−4	0.994	0.000	0.985	0.010	0.749	0.021	0.689
log(Start speed)	0.134	0.059	0.683	0.028	−0.043	0.000	0.103	0.572
log(End speed)	0.222	0.001	−0.374	0.192	0.377	0.000	0.520	0.002
log(zone)	−0.014	0.543	−0.014	0.539	−0.017	0.405	−0.020	0.581
log(Nasty)	−0.066	0.143	−0.065	0.048	−0.046	0.212	−0.095	0.163
Strike count	0.860	0.000	0.860	0.000	0.826	0.000	1.391	0.000
Inning	−0.063	0.071	−0.066	0.072	−0.058	0.035	−0.107	0.039
Outs	0.007	0.680	0.006	0.728	0.008	0.588	0.016	0.557
Batter hand	0.063	0.138	0.063	0.152	0.056	0.126	0.116	0.083
CONSTANT	−2.609	0.000	−2.442	0.000	−2.358	0.000	−4.829	0.001
$\widehat{\psi}$							0.052	0.000
$\widehat{\alpha}$			34370.714	0.433	−0.146	0.000		
$\widehat{\phi}$	−0.045	0.004	−0.046	0.000	0.110	0.000	0.074	0.001
$\ell_{\max}$	−3687.518		−3687.498		−3655.151		−3526.485	
MAE	0.671		0.671		0.670		0.647	
RMSE	0.850		0.850		0.853		0.847	

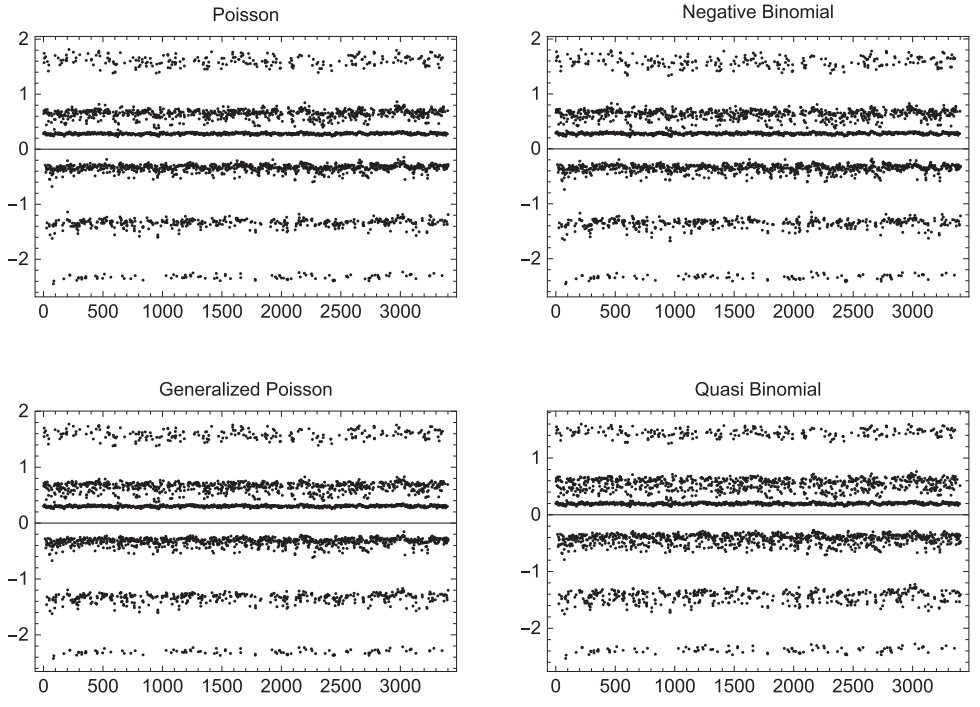

Figure 2. Raw residuals for the models considered. Baseball data.

Additionally, Figure 2 shows the raw residuals for the models considered, revealing again a greater dispersion of the Poisson, negative binomial and generalised Poisson distributions than of the QB distribution. Finally, Vuong test and the corresponding *p*-values (in

Table 10. Vuong tests comparing the QB model with the Poisson, NB and GP models for baseball data. The *p*-value is shown below.

	Homogeneous model		Inflated model	
	Without covariates	With covariates	Without covariates	With covariates
QB vs. Poisson	8.621	25.643	4.19	15.01
	< 0.001	< 0.001	< 0.001	< 0.001
QB vs. NB	9.85	25.434	4.19	15.09
	< 0.001	< 0.001	< 0.001	< 0.001
QB vs. GP	9.74	14.90	1.50	7.48
	< 0.001	< 0.001	0.134	< 0.001

brackets) were computed for this baseball data (see Table 10) concluding that the QB model achieves the best fit to the model.

5.2.4. Further development and interpretation

The model could be further developed to explicitly allow for covariates to model the inflation parameter ϕ. Now, we assume that this parameter is not constant for all observations but is modelled in a similar way to the mean parameter, depending on the covariates. This additional model can be specified by using (15) together with the usual log link for the scale parameter ϕ. That is, we consider $\phi(\boldsymbol{y_i}, \zeta) = \exp(\boldsymbol{y_i}^{\top}\zeta) \in (0, \infty)$, for $i = 1, \ldots, n$, where $\zeta = (\zeta_1, \ldots, \zeta_r)^{\top}$ is a vector with dimension r of unknown regression coefficients, different from those considered for θ_i. Clearly, both θ and ϕ, may be influenced by different characteristics and variables. For this reason, the explanatory variables used to model them may not be the same. The results obtained are shown in this case in Table 11. It is appreciated that all covariates are statistically significant for the inflation parameter with different signs in many of the cases compared to the case in which the dependent variable takes a value larger than zero. Then, the inclusion of this model seems to have an important effect on the dependence of the variable studied.

For the baseball data, by fixing the rest of covariates, the mean of the balls before the pitch is increased 1.021 times when the batter swings at the pitch as compared to not swing.

Table 11. Coefficient estimates and *p*-values for the zero inflated models with covariates also on the inflated parameter ϕ for baseball data.

	No inflated parameter		Inflated parameter	
Variable	Estimate	$\Pr > \|t\|$	Variable	$\Pr > \|t\|$
Batter number	0.011	0.088	−0.010	0.003
Swing	0.005	0.922	−5.669	0.000
log(Start speed)	0.067	0.697	−0.481	0.000
log(End speed)	0.481	0.001	−0.069	0.065
log(zone)	−0.023	0.508	−0.577	0.000
log(Nasty)	−0.102	0.115	−0.442	0.000
Strike count	1.300	0.000	−0.910	0.000
Inning	−0.104	0.047	−0.235	0.000
Outs	0.012	0.631	−0.800	0.000
Batter hand	0.113	0.061	−1.842	0.000
$\widehat{\psi}$	0.071	0.000		
CONSTANT	−4.466	0.000	−6.340	0.000
$\ell_{\max}$	−3531.917			

In a similar way, the mean of the balls before the pitch is 1.123 times for batter's stance right compared to batter's stance left. Analogously, the mean of the balls before the pitch is increased 4.019 times for each strike before the pitch and is also increased 1.619 for each mile per hour that is increased the speed crossing home plate.

6. Concluding remarks

In this study, we analyse the quasi binomial distribution: homogeneous and non-homogeneous, with and without inflation of zeros. Since this distribution is a generalisation of the classical binomial distribution, our findings can be considered a generalisation of the study by Hall [24].

The main advantage of the proposed model is that it takes into account that the support of the variable under study is bounded, a characteristic not present in the classical models based on the use of the Poisson, negative binomial and generalised Poisson distributions, among others. In practice, many of the variables under study have a limited support, as occurs for example in surveys of consumption in which the practitioner often requires respondents to report actual behaviour over a relatively short period of past time; and in biomedical research, where the practitioner may have to work with count data from dental epidemiological studies (for example of caries, in which case the variable is bounded by the number of teeth).

The proposed model has the advantage of its simplicity, since it does not incorporate any special function in its formulation. The normal equations that enable us to obtain the estimators of the parameters are simple, although not closed, but the elements of the Fisher information matrix that provide the standard errors are given in closed form. We believe the two models, whether inflated with zeros or non-inflated, with or without regression, provide a new and valuable perspective, one that can be used in many disciplines.

Acknowledgments

EGD also acknowledges the Departamento de Matemáticas, Facultad de Ciencias Básicas, Universidad de Antofagasta, Antofagasta (Chile) for their special support, as part of this paper was written while EGD was visiting the latter university in 2018.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

EGD was partially funded by grant ECO2013-47092 (Ministerio de Economía y Competitividad, Spain) and ECO2017-85577-P (Ministerio de Economía, Industria y Competitividad. Agencia Estatal de Investigación). ECO2017-85577-P, Emilio Gómez-Déniz is grateful for support received from MINEDUC-UA project, code ANT1755, since part of this paper was written while he was visiting the University of Antofagasta in 2019.

References

[1] T. Alanko and J. Duffy, *Compound binomial distributions for modelling consumption data*, The Statistician 45 (1996), pp. 269–286.

[2] D. Böhning, E. Dietz, P. Schlattmann, L. Mendonça, and U. Kirchner, *The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology*, J. R. Stat. Soc. A (Stat Soc) 162 (1999), pp. 195–209.
[3] J. van den Broek, *A score test for zero inflation in a Poisson distribution*, Biometrics 51 (1995), pp. 738–743.
[4] C. Brooks, *RATS Handbook to Accompany Introductory Econometrics for Finance*, Cambridge University Press, New York, 2009.
[5] C.G. Broyden, *A class of methods for solving nonlinear simultaneous equations*, Math. Comput. 19 (1965), pp. 577–593.
[6] C.G. Broyden, *Quasi-Newton methods and their application to function minimisation*, Math. Comput. 21 (1967), pp. 368–381.
[7] A.C. Cameron and P.K. Trivedi, *Econometric models based on count data.comparisons and applications of some estimators and tests*, J. Appl. Econom. 1 (1986), pp. 29–53.
[8] C. Cameron and P. Trivedi, *Regression Analysis of Count Data*, Cambridge University Press, New York, 1998.
[9] M. Campbell, D. Machin, and C. D'Arcangues, *Copying with extra Poisson variability in the analysis of factor influencing vaginal ring expulsion*, Stat. Med. 10 (1991), pp. 241–251.
[10] Ch.A. Charalambides, *Abel series distributions with applications to fluctuations of sample functions of stochastic processes*, Commun. Stat. Theory Methods 19 (1990), pp. 317–335.
[11] A.C. Cohen, *A note on certain discrete mixed distributions*, Biometrics 22 (1966), pp. 566–572.
[12] P.C. Consul, *A simple urn model dependent upon predetermined strategy*, Sankhyā: Ind. J. Stat. B 36 (1974), pp. 391–399.
[13] P.C. Consul, *On some properties and applications of quasi-binomial distribution*, Commun. Stat. Theory Methods 19 (1990), pp. 477–504.
[14] P.C. Consul and F. Famoye, *Generalized Poisson regression model*, Commun. Stat. Theory Methods 21 (1992), pp. 89–109.
[15] P.C. Consul and G.C. Jain, *A generalization of the Poisson distribution*, Technometrics 15 (1973), pp. 791–799.
[16] D. Cox and D. Hinkley, *Theoretical Statistics*, Chapman and Hall, London, 1974.
[17] C. Demètrio, *Letter to the editor: Copying with extra poisson variability in the analysis of factor influencing vaginal ring expulsion*, Stat. Med. 13 (1994), pp. 873–876.
[18] M. Denuit, X. Marèchal, S. Pitrebois, and J.-F. Walhin, *Actuarial Modelling of Claim Counts Risk Classification, Credibility and Bonus-Malus Systems*, John Wiley & Sons, Great Britain, 2009.
[19] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall/CRC, New York, NY, USA, 1993.
[20] F. Famoye and K. Singh, *Zero-inflated generalized Poisson regression model with an application to domestic violence data*, J. Data Sci. 4 (2006), pp. 117–130.
[21] R. Fletcher and M. Powell, *A rapidly convergent descent method for minimisation*, Comput. J. 6 (1963), pp. 163–168.
[22] P.L. Gupta, R.C. Gupta, and R.C. Tripathi, *Score test for zero inflated generalized poisson regression model*, Commun. Stat. Theory Methods 33 (2005), pp. 47–64.
[23] S. Gurmu and P. Trivedi, *Excess of zeros in count model for recreational trips*, J. Bus. Econ. Stat. 14 (1996), pp. 469–477.
[24] D.B. Hall, *Zero-inflated Poisson and binomial regression with random effects: A case study*, Biometrics 56 (2000), pp. 1030–1039.
[25] A. Hassan and M. Ahmed, *A probabilistic dose dependent model and survival analysis*, J. Stat. Theory Appl. 11 (2012), pp. 197–208.
[26] D. Lambert, *Zero-inflated Poisson regression, with an application to defects in manufacturing*, Technometrics 34(1) (1992), pp. 1–14.
[27] J.D. Lewsey and W.M. Thomson, *The utility of the zero-inflated poisson and zero-inflated negative binomial models: A case study of cross-sectional and longitudinal DMF data examining the effect of socio-economic status*, Community Dent Oral Epidemiol. 32 (2004), pp. 183–189.
[28] D. Lord, S.P. Washington, and J.N. Ivan, *Poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory*, Accid. Anal. Prev. 37 (2005), pp. 35–46.

[29] H. Ruskeepaa, *Mathematica Navigator. Mathematics, Statistics, and Graphics*, 3rd ed., Academic Press, USA, 2009.
[30] L. Shenton, *Quasibinomial distributions*, Encyclopedia of Statistical Sciences, 2006.
[31] Q.H. Vuong, *Likelihood ratio tests for model selection and non-nested hypotheses*, Econometrica 57 (1989), pp. 307–333.
[32] R. Wilcox, *Fundamentals of Modern Statistical Methods. Substantially Improving Power and Accuracy*, 2nd ed., Springer, 2010.
[33] R. Winkelmann, *Econometric Analysis of Count Data*, Springer, Berlin Heidelberg, Germany, 2003.

Appendices

Appendix 1

The following formulas can be used to compute the information matrix in the ZIQB distribution without covariates.

$$E\left[\frac{X}{(p+X\psi)^2}\right] = mp(p+\psi)^{-2} + m_{(2)}p^2(p+\psi)^{-1}(p+2\psi)^{-1},$$

$$E[(p+X\psi)^{-2}] = p^{-2} - m\psi[(p+\psi)^{-1}p^{-1} + (p+\psi)^{-2}] + m_{(2)}\psi^2(p+\psi)^{-1}(p+2\psi)^{-1},$$

$$E\left[\frac{m-X}{(1-p-X\psi)^2}\right] = m(1+\psi-m\psi)(1-p-m\psi+\psi)^{-1},$$

$$E\left[\frac{(X-1)X}{(p+X\psi)^2}\right] = m_{(2)}p(p+2\psi)^{-1},$$

$$E\left[\frac{(m-X)X}{(1-p-X\psi)^2}\right] = m(m-1)p(1-p-m\psi+\psi)^{-1},$$

$$E\left[\frac{(X-1)X^2}{(p+X\psi)^2}\right] = \frac{2pm_{(2)}}{p+2\psi} + p\sum_{k=0}^{m-3}\psi^k m_{(k+3)},$$

$$E\left[\frac{(m-X)X^2}{(1-p-x\psi)^2}\right] = m(m-1)^2p(1-p-m\psi+\psi)^{-1} - p\sum_{k=0}^{m-3}\psi^k m_{(k+3)}.$$

Appendix 2

This simple code is written in R and can be used to estimate the ZIQB model with covariates (example with only two covariates and an intercept is provided).

```
setwd("C:/Users/Usuario/Desktop")
initialValues=read.delim ("DOCTORAUST.txt",header=FALSE)
rm(list=ls())
library(maxLik)
gender=initialValues[,1]
age=initialValues[,2]
actdays=initialValues[,5]
n <- length(actdays)
constant <- replicate(n,1)
m = 14
theta <- function(b0,b1,b2){m*exp(b0*constant+b1*gender+b2*age)/
```

```
      (1+exp(b0*constant+b1*gender+b2*age))}
k <- c(0,1,2,3,4,5,6,7,8,9,10,11,12,13)
p <- function(theta,fi){(theta/m)*(sum((fi^k)*gamma(m)/gamma(m-k)))}
Indicador=ifelse( actdays==0, 1, 0)
dziqbcov =function(y,I,phi,fi,b0,b1,b2){(((1/(1+phi))*
        (phi+(1-p(theta(b0,b1,b2),fi))^(m)))^I)*((1/(1+phi))*
         (factorial(m)/(factorial(y)
logLikFunqbcov <- function(param) {
  phi <- param[1]
  fi <- param[2]
  b0 <- param[3]
  b1 <- param[4]
  b2 <- param[5]
  log(dziqbcov(actdays,Indicador, phi, fi, b0, b1, b2))}
mle <- maxLik(logLik = logLikFunqbcov,start=c(phi =5, fi = 0.001,
 b0 = -1.5, b1 = 0.02, b2 = 0.1),method = "BHHH")
summary(mle)
```